Residual Load Forecasting: Kaggle Challenge

Chloe Duane, Diana Salasplata, Mahdieh Sheikh Rezaei, Roberto Verdezoto

University of Houston's Data Science for Energy Transition Program
Sam Houston State University, Huntsville, Texas

1. PROBLEM STATEMENT

Participants of the challenge were asked to develop an algorithm that tries to forecast the residual load. Official task was to predict the residual_load for the times contained in the test.csv file.

The residual load is understood as the remaining energy demand that is still left to be made for a country after they have used their entire energy capacity.

It is important for energy suppliers to know how much residual load they are going to need to have their energy needs complete.

2. DATA SOURCE

Data was provided by the Kaggle competition organizers.

The data provided was a test dataset **test.csv** and a training dataset **train.csv**

3. PARAMETERS

Name	Description	Contained in set
Time	Timestamp	Train; Test
P	Amount of power generated by the photovoltaic system	Train
Gb(i)	Direct in-plane irradiance. The fraction of the solar radiation that directly reaches the ground. It is only available if clouds do not block the sun.	Train; Test
Gd(i)	Diffuse in-plane irradiance. It is the fraction of solar radiation that reaches the ground after being reflected or scattered by the atmosphere. Its a also available if clouds block the sun.	Train; Test
H_sun	Sun height (elevation)	Train; Test
T2m	Air temperature at 2m	Train; Test
WS10m	Wind speed at 10m	Train; Test
load	Amount of power required by the plant at this timestep.	Train
residual_load	residual_load = load-P. If this value is positive: The amount of power that has to be taken from the grid. If this value is negative: The amount of energy which is fed into the electricity grid.	Train
dataset_id	Train and Test data with the same dataset_id belong together	Train; Test

4. FORECASTING MODEL APPROACHES

- Along the different ways of approaching a forecast model, members took three different tracks: time series, linear regression model random forest model for forecasting. Data was treated as a categorical variable.
- For the time series approach, the SARIMAX algorithm was used. SARIMAX was a good approach because the model takes into consideration the seasonality and external factors: How energy usage and sun angles will be different at different points in time.
- For the Random Forest Regressor, it's used for regression and in our case to predict 'load' and 'P". 100 decision trees were used to predict load. With how big the data set is, 100 was the most solid option since the more trees you generate the more time it will take to compile, effecting the overall V1 score. Then the residual load is calculated by the equation residual_load='load'-'P', with the previously predicted variables.

5. RESULTS

```
Call:
lm(formula = residual_load ~ time + Gb.i. + Gd.i. + H_sun + T2m +
    WS10m + dataset_id, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-451.23  -24.17   -4.57   16.97  508.43

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  2.395e+03  1.541e+02   15.544  < 2e-16 ***
time        -1.472e-06  1.015e-07  -14.506  < 2e-16 ***
Gb.i.       -5.271e-01  1.838e-03 -286.835  < 2e-16 ***
Gd.i.       -4.687e-01  5.837e-03  -80.297  < 2e-16 ***
H_sun        6.081e-02  3.798e-02    1.601  0.10932
T2m         -9.486e-01  4.715e-02  -20.118  < 2e-16 ***
WS10m        7.901e-01  1.674e-01    4.721 2.35e-06 ***
dataset_id2 -2.634e+01  1.451e+00  -18.147  < 2e-16 ***
dataset_id3 -5.985e+00  2.109e+00   -2.838  0.00454 **
dataset_id4  2.782e+01  2.666e+00   10.432  < 2e-16 ***
dataset_id5  4.116e+01  3.416e+00   12.047  < 2e-16 ***
dataset_id6  2.153e+01  4.261e+00    5.053 4.36e-07 ***
dataset_id7  3.569e+01  5.082e+00    7.022 2.20e-12 ***
dataset_id8  7.523e+01  5.810e+00   12.948  < 2e-16 ***
dataset_id9  7.032e+01  6.606e+00   10.644  < 2e-16 ***
dataset_id10 5.586e+01  7.456e+00    7.491 6.88e-14 ***
dataset_id11 8.368e+01  8.283e+00   10.102  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63.19 on 88695 degrees of freedom
Multiple R-squared:  0.8074,    Adjusted R-squared:  0.8073
F-statistic: 2.323e+04 on 16 and 88695 DF,  p-value: < 2.2e-16
```

Fig. 1: Summary report for the linear regression model done in R studio. $R^2 = 0.8073$

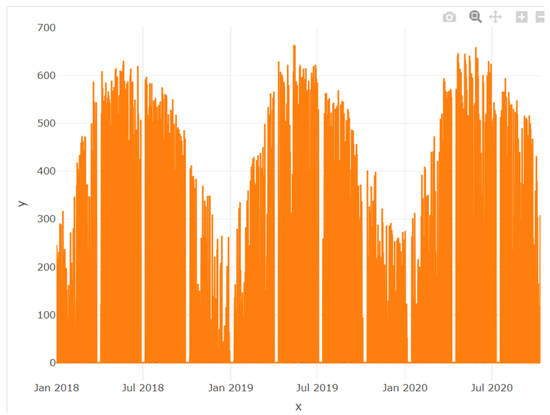

Fig. 3: Trend of P

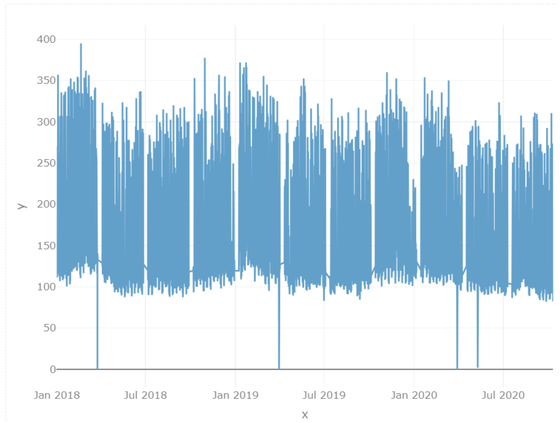

Fig. 2: Trend of load

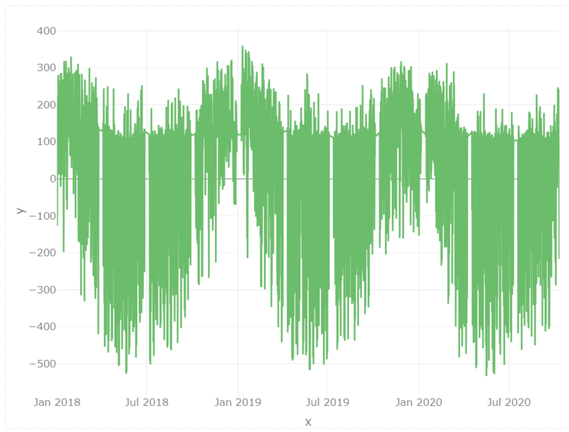

Fig. 4: Time Trend of residual load = load - P

5. RESULTS CONT.

- Overall, the two approaches that gave the best results for forecasting the residual load were the linear regression and the random forest. They got a grade of 62 and 62.6 respectively in Kaggle.
- Unfortunately the SARIMAX approach was not able to run and kept timing out as it was tried in the software.

6. CONCLUSIONS

- Most of the variables were significant for the analysis. And both the time series and random forest approach got a For a preliminary model is a perfect starting point.
- Some of the challenges presented was the formatting of the time variable. When tried to do the time series model with R studio, the time field had to be changed to Epoch Seconds and brought back again to regular format.
- Due to time constraints, there was no opportunity to try other ways of adjusting the linear model.
- Nevertheless this represent as an excellent starting point for making a forecast for residual load. Since forecasting this feature will be getting more complex in the future with the additional of new ways of energy.
- Having a good forecast model for the residual load of a grid can help a country greatly in their energy transition process.

REFERENCES

Tobias Rohrer. (2023). Energy Forecasting Data Challenge Public. Kaggle. https://kaggle.com/competitions/energy-forecasting-data-challenge-public